Uso del DERIVE en el cálculo de la matriz de varianza y covarianza y otros factores

Eliseo Martínez

1. Introducción

No hay duda que el software DERIVE no es el indicado para realizar métodos estadísticos multivariantes, como lo son el STATGRAPHICS, MINITAB o SPSS. No obstante, trabajar con el DERIVE con los primeros cálculos necesarios en métodos multivariantes puede ser didácticamente útil en la comprensión de los primeros conceptos.

En el presente trabajo daremos indicaciones sobre el uso de un pequeño programa realizado en DERIVE, que nos permitirá hacer sencillos cálculos iniciales en el análisis multivariante a partir de una matriz de datos $\mathbf{X}$ de orden $n \times p$.

El orden de desarrollo será el siguiente:

- Daremos instrucciones para leer la matriz de datos $\mathbf{X}$ mediante el DERIVE, para esto la base de datos de esta matriz debe estar en extensión **.dat**
- Definida la matriz de datos $\mathbf{X}$ en el software DERIVE, daremos la sentencia que nos permite calcular el vector de medias, que llamaremos $\overline{\mathbf{x}}$ de tal forma que cada componente es la media de cada una de las columnas de la matriz $\mathbf{X}$.
- Una vez obtenido el vector de medias, formaremos la matriz centrada $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^t$
- De la matriz centrada realizaremos el cálculo de la matriz de varianza y covarianza, donde
$$\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$$
- Con la matriz S vamos a definir la distancia de Mahalanobis, que es
$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)^t\, \mathbf{S}^{-1}\, (\mathbf{x}_i - \mathbf{x}_j)\right]^{\frac{1}{2}}$$
entendiendo que los vectores $\mathbf{x}_i$ y $\mathbf{x}_j$ son las filas i y j de la matriz de datos $\mathbf{X}$.
- Finalmente se realiza el cálculo de la matriz de correlación $\mathbf{R}$ a través de los coeficientes lineales de Pearson entre los vectores columnas de la matriz $\mathbf{X}$.

2. El programa multiv.mth[1]

Cargue este fichero en el software DERIVE. Aparecerán las siguientes instrucciones que

[1] Este programa lo puede bajar desde la red Internet en la ubicación http://www.uantof.cl/facultades/csbasicas/matematicas/academicos/emartinez/magister/multiv.mth

reproduciremos

#1 X es una matriz de datos de n filas y m columnas
#2 Defina como X : = a la matriz de datos
#3 X : =
#4 n : = DIM(X)
#5 p : = DIM(X‘)
#6 a (i, j) : = ELEMENT(X, i, j)
#7 med(j) : = (1 / n)SUM (a(i , j), i, 1, n)
#8 Cálculo del vector de medias
#9 media : = VECTOR(VECTOR(med(j), j, 1, p), k, 1, 1)
#10 Cálculo de la matriz de datos centrada
#11 uno : = VECTOR(VECTOR(1, i, 1, n), j, 1, 1)
#12 Xcen := X - uno‘· media
#13 Calculo de la matriz de varianza y covarianzas S
#14 S : = (1/(n-1)) Xcen‘· Xcen
#15 Calculo de la distancia de Mahalanobis d(i, j)
#16 Fila(i) : = VECTOR(VECTOR(a(i, j), j, 1, p), h, 1, 1)
#17 d(i, j) : = (((Fila(i) - Fila(j))S^{-1}(Fila(i) - Fila(j))‘))$^{1/2}$
#18 Cálculo de la matriz de correlación R
#19 b(i, j) := ELEMENT(S, i, j)
#20 r(i, j): = b(i, j) / ($\sqrt{\text{b(i, i)}}\sqrt{\text{b(j, j)}}$
#21 R : = VECTOR(VECTOR(r(i, j), i, 1, p), j, 1, p)

3. La matriz de datos X

Vamos a suponber que la matriz de datos está en extensión **.dat** o en extensión **.txt,** ,luego se sigue la secuencia **Archivo** $>$ **Leer** $>$ **Datos.** Esta matriz de datos aparecerá en la pantalla de trabajo del DERIVE, entonces inmediatamente hay que posicionarse en la matriz con el video reverso, y escribir en la linea editora X : =, seguido de la recla F3 que es la que copia lo que está marcado en el video reverso. Hecha esta operación podemos ir haciendo los cálculos paulatinamente.

Las instrucciones #4 y #5 nos permiten saber las dimensiones filas y columnas de la matriz. Debemos hacer notar que la instrucción Dim(A), siendo A una matriz, entrega el número de filas, de modo que para saber el número de columnas se debe trasponer la matriz A, y esta operación se hace mediante A´, de modo que Dim(X´) nos entrega el número de columnas de la matriz $\mathbf{X}$. La identificación de las filas y columnas nos permitirá almacenarlas en las variables n y p, puesto que las ocuparemos para cálculos posteriores. Finalmente guardamos todas las entradas de $\mathbf{X}$ definiendo las variables a(i, j) como lo indica la instrucción #6.

4. El vector de medias

Habiendo definido las entradas a(i, j) de la matriz $\mathbf{X}$, calculamos las p medias mediante **med(j)** en la instrucción #7. Ahora bien, hemos dado la instrucción matricial en #9 para definir el vector de medias, que hemos llamado **media**, como una matriz de $1 \times p$ más que un vector de dimensión p. Y la diferencia entre vector de dimensión p y matriz de dimensión $1 \times p$ es fundamental, puesto que entre vectores los únicos productos que se permiten, en el DERIVE, son el escalar y el producto cruz, y no se puede formar entre dos vectores de dimensión p y dimensión n, respectivamente, una matriz de tamaño $p \times n$ con la sola traspuesta del primer vector.

5. La matriz de datos centrada

La matriz de datos centrada está definida como $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^t$, de manera que necesitamos definir el vector de dimensión n (en rigor, una matriz de dimensión $n \times 1$) cuyas entradas sean todos "unos" tal como lo establece la instrucción #11, y que hemos definido como **uno**; y la instrucción #12 es el cálculo de la matriz centrada teniendo en cuenta que en la fórmula $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}^t$ se consideran los vectores como columnas, al contrario que en el DERIVE, por eso la diferencia en la aplicación de la traspuesta. En rigor Xcen := X - uno‘· media, donde Xcen = $\widetilde{\mathbf{X}}$.

El cálculo de la matriz de varianzas y covarianzas es practicamente directo, esto es el producto matricial de las matrices centradas y dividido por el factor $(n-1)$, y se realiza mediante la instrucción #14.

6. La distancia de Mahalanobis

Esta distancia tiene sentido definirla entre los vectores filas de la matriz de datos $\mathbf{X}$, de modo que es necesario rescatar las filas de la matriz $\mathbf{X}$, que en rigor son vectores filas de dimensión p, como matrices de dimensión $1 \times p$, y tendremos n matrices de este tipo, y cada una de ellas las hemos denotado por **Fila(i)** según la instrucción #16. Con esta notación la definición de la distancia de Mahalanobis es inmediata, mediante la instrucción #17.

7. La matriz de correlación R

Esta matriz es obtenida manipulando las entradas de la matriz $\mathbf{S}$. Considerando que los elementos de la diagonal de $\mathbf{S}$ son las varianzas y los elementos fuera de la diagonal las covarianzas, de modo que ellas definen el llamado coeficiente de correlación de Pearson, y que así lo establece la instrucción #20, y de esta forma definir la matriz de correlación $\mathbf{R}$.